

# On The Discovery of Social Roles in Large Scale Social Systems

Derek Doran

Department of Computer Science & Engineering

Kno.e.sis Research Center

Wright State University, Dayton, OH

derek.doran@wright.edu

## Abstract

*The social role of a participant in a social system is a label conceptualizing the circumstances under which she interacts within it. They may be used as a theoretical tool that explains why and how users participate in an online social system. Social role analysis also serves practical purposes, such as reducing the structure of complex systems to relationships among roles rather than alters, and enabling a comparison of social systems that emerge in similar contexts. This article presents a data-driven approach for the discovery of social roles in large scale social systems. Motivated by an analysis of the present art, the method discovers roles by the conditional triad censuses of user ego-networks, which is a promising tool because they capture the degree to which basic social forces push upon a user to interact with others. Clusters of censuses, inferred from samples of large scale network carefully chosen to preserve local structural properties, define the social roles. The promise of the method is demonstrated by discussing and discovering the roles that emerge in both Facebook and Wikipedia. The article concludes with a discussion of the challenges and future opportunities in the discovery of social roles in large social systems.*

## 1 Introduction and Motivation

Why do people choose to participate and interact with others in a social system? This basic question lies at the heart of many sociological studies that examine the nature of interactions in a community. The question is theoretically associated with the *social roles* of community members, which is defined as a qualitative description

capturing the circumstances and reasons under which they choose to interact with others. The concept of a social role is fundamentally based on the notion of a user's *position* within a social network [39, 62]. For example, with whom and how one decides to connect to others in a community is associated with how they are perceived by others [13], the power they hold [22], and their ability to spread information and influence others [98]. As a concrete illustration, consider a network of interactions among workers in a corporate office. Some workers have the social role “manager” as defined by who they are connected to socially: “managers” are responsible for the work of the “team members” he leads and report to an “executive”. Any person in the corporate office network in a similar position, even if they report to a different executive and managers a different team, is still perceived to have the social role “manager”.

Extracting and understanding the social roles of a social system carries theoretical and practical importance. Theoretically, an analyst may integrate the social roles discovered in a social setting and the context of these interactions to formulate a thesis about the reasons why and how people interact within the system. For example, consider the typical interactions that may occur within a generic corporate office as well as the connotations of being labeled a “manager”. Analysts could infer that “managers” interact with “team members” based on the initiatives and projects assigned to them by “executives”. They may be required to balance the demands placed on them by executives along with the needs of the team, and serve as a broker that filters information from corporate leaders to others in the organization. Practically, the delineation of users by their social role facilitates the interpretation of complex social systems by simplifying their structure from connections among users to between roles [8, 78, 95]. It also enables meaningful studies of communities across time and context (e.g., different types corporate offices) by comparing the structure of interactions between roles that are common among them. For example, meta-analysis of the social roles and the interactions among them across different groups can help designers create effective physical and digital spaces for communities and organizations to grow within [44]. Social role analysis is also useful to identify the types of users that may become influential [40], and even reveal latent social structures within the systems [58].

This article presents a new method to discover the social roles that exist in large scale online social systems. The methodology is motivated by an analysis of the present art, which either: (i) requires an analyst to presume the existence of roles beforehand; and/or (ii) mines the roles using features about the users and the structure of the system that may not have a basis in social theory. The approach discovers social roles by clustering users by their *conditional triad census*, which is a vector capturing the types and orientations of three way relationships their ego-network is composed of. The method is applied to a network of interactions from an online social network (Facebook) and a collaborative editing platform (Wikipedia). An analysis of the quality of the resulting

clusters and the ego-network structure of prototypical users demonstrate the utility of the proposed method. The article concludes with a discussion about the many opportunities and challenges for future research in social role discovery for large scale social systems.

This article is organized as follows: Section 2 reviews and assesses existing methods for social role discovery in large scale social systems. Section 3 introduces the concept of a conditional triad census and the proposed methodology. Section 4 analyzes the structure of the social roles mined from two large scale online social systems. Important challenges and opportunities that remain in the analysis of social roles in large scale systems are presented in Section 5. Concluding remarks are offered in Section 6.

## 2 Discovering Social Roles

Present methods to discover social roles in social systems may be classified into three types: (i) methods that define roles by notions of equivalence; (ii) methods that require the assertion of the roles existing in the system prior to analysis; and (iii) methods that define roles based on patterns among user attributes and system interactions. This section provides an overview of each type and their applicability to discover social roles in large scale systems.

### 2.1 Equivalence based role discovery

Longstanding methods to identify social roles are based on finding users who are in “equivalent” positions [95, 9, 10, 11], which may be defined in one of three ways. Given an undirected network  $G = (V, E)$  of users  $V$  connected by a set of relations  $E$ , *structural equivalence* requires two users  $i$  and  $j$  to be connected to be exactly the same set of others. In other words, for every relationship  $(i, x) \in E$  that exists, the relation  $(j, x)$  must also exist. Under this definition, a user’s social role is precisely defined by the people that she is connected to. This strict definition may not be useful in many settings because it is impossible for two users whose distance is greater than two in a network to fall under the same role. For example, two “managers” in an office that report to a common “executive” but have difference sets of subordinates are not structurally equivalent and would therefore not be classified under the same role.

*Isomorphic equivalence* offers a broader definition of equivalent network positions. An isomorphism among two users in a network exist if there is a mapping  $\pi : E(a) \rightarrow E(b)$  where  $E(a)$  is the set of relationships held by user  $a$  such that for every pair of users  $a, b \in E$ , we have  $(a, b) \in E(a)$  if and only if  $(\pi(a), \pi(b)) \in E(b)$ . In other words, users  $a$  and  $b$  must have isomorphic *ego-networks*, which is a tuple  $(V_e, E_e)$  where  $V_e$  is the set of all users in the  $2^{nd}$  degree neighborhood of a user and  $E_e$  represents the directed relationships that bind the users in  $V_e$  together. This suggests that one could simply switch the location of user  $a$  and  $b$  and their connectivity to others

without disturbing the overall structure of the network. Practically, two “managers” in a network that report to an “executive” and lead the same number of “team members” would be isomorphically equivalent if the connectivity among the “team members” of the two “managers” were isomorphic. This equivalence definition thus captures a more intuitive notion for ascribing a user’s role in a social system. A still broader class is *regular equivalence*, which requires the role of the alters of two users to be identical. Specifically, if  $\mathcal{R}(x)$  is a function that assigns a user  $x$  to a role, we say users  $a$  and  $b$  are regularly equivalent if  $\mathcal{R}(a) = \mathcal{R}(b)$  and if every user  $n$  in the ego-network  $N(a)$  of  $a$  can be mapped to a user  $m$  in the ego-network of  $b$  such that  $\mathcal{R}(n) = \mathcal{R}(m)$ . For example, “managers” would be regularly equivalent so long as they both connect to “executives” and “team members”. Isomorphic and regular equivalences may be identified by performing a blockmodeling over the adjacency matrix of a social system [89].

Notions of structural, isomorphic, and regular equivalence are decades old theories that have been instrumental in many social network analyses [82, 25, 23, 91, 94, 30]. More recent work have used these notions to study international relationships across institutions [69], firms [73], governments [101, 53], and to study peer influences [36]. Isomorphic equivalence has been applied to hospitals within referral networks [50] to discover closed communities of health services and hospitals that carry identical areas of expertise. They are also employed in the study of citation networks [85] to identify researchers within an organization that perform similar research and offer similar domain expertise. Regular equivalences have been studied in networks of relations among gang members in urban settings [75] and of relations among cities across the world [2].

## 2.2 Implied role discovery

In implied role analysis, a researcher defines the set of social roles users of a social system are expected to exhibit before any data or structural analysis commences. It is a qualitative, iterative process that generally follows the workflow of Figure 1. Based on at-hand information about a social system, roles are first defined based on the subset of functionality allowed by the system that the user may perform. For example, consider an online forum where users may decide to browse conversations but never post, or can become an administrator that edits and controls the behavior of others in the system. An analyst may therefore first define the social roles *lurker* (one who never posts), *moderator* (one who controls behavior), and *poster* (one who contributes to conversations). With these roles assumed to exist, the analyst studies the actions of users and their relations with others. The initial definitions of the social roles are then iteratively refined as evidence from the social system is collected.

Implied role analysis is useful when a social system is well understood, highly structured, and if the analyst wishes to understand the interactions among users on the basis of the kinds of operations they perform. For

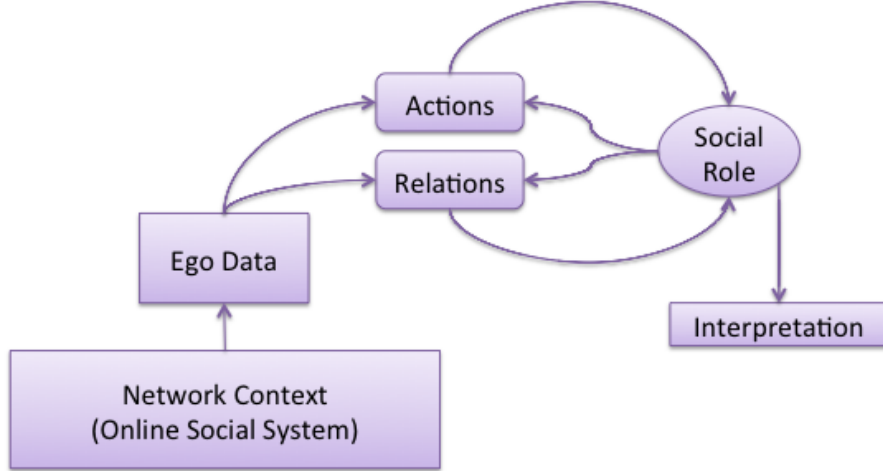


Figure 1: Workflow for implied-role analysis

example, Nölker *et al.* tapped into their experiences with online bulletin board systems to predefine members of a Usenet group into the roles *leader*, *motivator*, and *chatter* [71]. They identified the behavioral attributes that are indicative of each role, and labeled users exhibiting such behaviors in a log of the group’s activity. Golder *et al.* also studied Usenet groups but proposed a different taxonomy of roles that include *celebrities*, *ranters*, *lurkers*, *trolls*, and *newbies* [41]. They sifted through conversations across different Usenet groups to study behaviors associated with each role. Gliwa *et al.* examined collections of online bloggers and defined the roles *selfish influential user*, *social influential user*, *selfish influential blogger*, *social influential blogger*, *influential commentator*, *standard commentator*, *not active*, and *standard blogger* [40]. Welser *et al.* defined four roles for Wikipedia users, namely *substantive experts*, *technical editors*, *counter vandalism*, and *social networkers* [92]. They subsequently searched for patterns about how users contribute and interact with others in order to classify the users falling in each role.

### 2.3 Data-driven role discovery

A third type of approach is to infer social roles by the features of a dataset without pre-defining the roles that exist. These data-driven approaches, whose workflow is summarized in Figure 2, generally considers features about users and the structure of their ego-networks in an unsupervised machine learning algorithm. Social roles are defined as the groups the algorithm places users into based on the similarity of these features. Studies that apply unsupervised learners for social role discovery vary in sophistication. For example, Hautz *et al.* categorized users in an online community of jewelry designers by mapping whether their out- and in-degree distributions and frequency of interactions to “low” or “high” levels [44]. Zhu *et al.* use *k*-means clustering to identify user roles in a network of phone calls based on similar calling behaviors, ego-network clustering coefficients, and mean geodesic

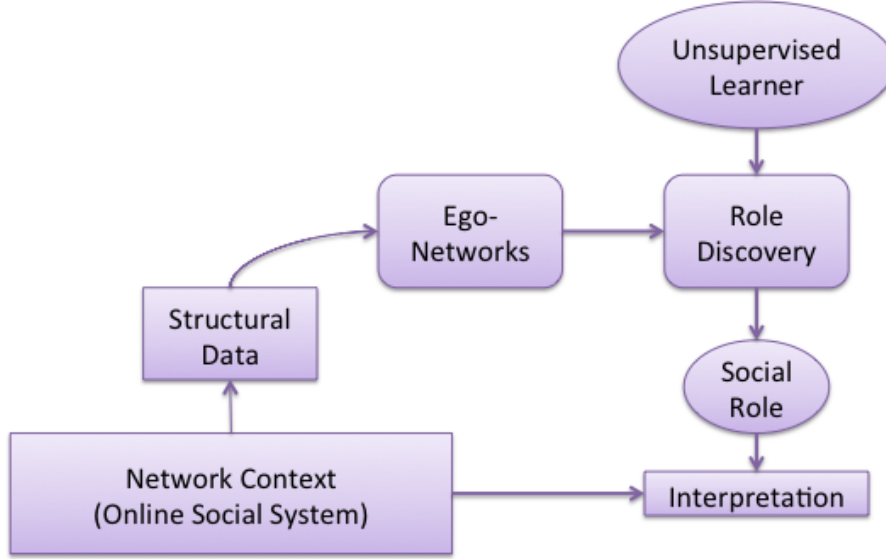


Figure 2: Workflow for data-driven role analysis

distances between users [102]. Chan *et al.* discover roles by agglomerative hierarchical clustering with over fifty behavioral and structural features of users' across the post/reply network of many online forums [18]. White *et al.* use a mixed membership probabilistic model to identify roles across online forums using behavioral features and found a number of possible assignments of users into groups [93]. Rowe *et al.* use behavioral ontologies and semantic rules to automatically group online forum users into roles based on the content of their posts [77]. Although data-driven approaches define similarity based on the structural features of ego-networks, this class of methods is not an approximation of equivalence based role discovery. This is because data-driven methods may search for the similarity of two users based on many feature types that are not structural, including their personal attributes, their behaviors on the social system, and the content of their interactions with others.

## 2.4 Comparative analysis

The recent availability of data about very large scale social systems, typically collected from online social networks (Facebook; Google+), social media (Twitter; Tumblr), and innovative information exchanges (Wikipedia; StackExchange) enables the study of the social roles of users in systems that have a world-wide reach. The massive scale of these systems necessitates the need to evaluate current approaches for discovering social roles, so that the most effective type given their size can be identified.

Equivalence based role discovery comprises a number of well-studied, longstanding methods that has deep roots in sociological theory. Unfortunately, it may be infeasible to precisely identify users falling into isomorphic

or regular equivalence classes within large scale social systems. This is because the problem of finding isomorphic ego-networks is closely aligned to searching for all motifs of arbitrary size within the network, and the problem of identifying regularly equivalent positions is related to searching for a  $k$ -coloring of  $G$ , with  $k$  unknown a priori (both are NP-hard problems [52]). Researchers still interested in identifying these equivalences in large systems must resort to numerical approximations based on quantitative notions of structural similarity between two users that may be difficult to apply and analyze in practice [70, 32, 49]. Thus, despite the rich theory they are grounded within, technical challenges bar its adequate adaptation for large scale social systems.

Implied role analyses carry fewer technical challenges. This is because the most difficult aspect - identifying the roles that exist - are predefined by an analyst before trends in the data are considered. However, implied role analyses runs the risk of using noisy signals in the data that appear by chance as evidence for the roles they have predefined. Furthermore, it is possible for separate analysts to define completely different sets of social roles for the same system, which may confuse or conflict each other. For example, Nolker *et al.* places Usenet members into *leader*, *motivator*, and *chatter* roles [71]. Are these roles compatible with the alternative set of *celebrities*, *ranters*, *lurkers*, *trolls*, and *newbie* roles proposed by Golder *et al.* for the same system [41]? It is unclear if one set of roles is more suitable than the other, or if the cross-product of the two types of roles (e.g. *leader-celebrity* or *chatter-lurker*) is also a valid set of roles. Furthermore, the implied roles tend to speak to the functionality or actions that users of the social system undertake instead of reflecting the reasons why they participate in the system and the way they are structurally embedded within it. Thus, although there are fewer technical challenges to run implied role analysis over large scale social systems, the resulting roles may have a weak relationship to sociological theory.

Data-driven social role analysis may be a promising type of approach for the discovery of social roles in large-scale social systems. This is because modern day “big data” technologies enable the collection of incredible amounts of information about each user, their connections with others in the social system, and the details or the content of their interactions. Instead of assuming that specific kinds of social roles in the system must exist, data-driven analyses apply data mining algorithms or learn data models from which the social roles of the system emerge. Such approaches let the data inform the analyst what social roles exist, rather than require a definition of the roles before studying the data. Fortunately, recent big data systems and methods research enable the rapid mining and building of data models from large social systems. For example, Zhang *et al.* tackle computations over real-world and virtual social interaction data by performing Tucker decompositions of a tensor representation of the interactions [99]. A distributed learning algorithm based on the MapReduce proposed by Tang *et al.* efficiently identifies the influencers and experts latent within large social systems [84]. Cambria *et al.* use a comparative anal-

ysis of the performance of multiple natural language processing algorithms to find patterns in the content of social interactions [16]. Giannakis *et al.* present a series of articles that describe how sensor signal processing algorithms may be adapted to operate over big and social data sets [38]. Malcom *et al.* even developed a uniform programming interface so that non-experts can utilize state-of-the-art big data technologies [63] for social role analysis. However, the relationship of such analyses with longstanding social theory varies considerably. This is because while some data mining algorithms and models encode aspects of social theory in their technical development, others were given no consideration to these theories in their development or make assumptions that are incompatible with past social science research for sake of model tractability. Furthermore, algorithms and models for social role analysis may use features that do not reflect aspects of social forces that drive users to embed themselves in the network in a specific way [18].

### 3 Triad-based Social Role Extraction

In this section, a new data-driven approach for extracting social roles from large social systems is introduced<sup>1</sup>. Based on the discussion in the previous section, it only considers features that have a grounding in social theory, namely the *conditional triads* that compose each user’s ego-network. After network sampling and dimensionality reduction, *k*-means clustering is applied to the vectors to identify social roles. Ego-networks falling closest to the centroid of each cluster is interpreted for role analysis. This section describes what conditional triads are, the triad-based representation of an ego-network, the social systems used to illustrate the methodology, and the role extraction process.

#### 3.1 Conditional Triad Census

In social network analysis, a *triad* is a group of three individuals and the pairwise interactions among them [79]. They are the smallest sociological unit from which the dynamics of a multi-person relationship can be observed, and hence, are considered to be the atomic unit of a social network [33, 90, 24]. For example, third actors may act as a moderating force that can resolve conflicts among two others [13]. They may also sabotage an existing relationship or induce a feeling of unwelcomeness to a specific alter [7]. Such observations have been used to develop theories that associate the configuration of a triad to specific underlying effects that promote specific kinds of social interactions [46, 6].

Figure 3 captures the 36 different ways an individual (white) can be oriented towards two alters (blue) within a

---

<sup>1</sup>Parts of this method were presented at the First Workshop on Interaction and Exchange in Social Media at the 2014 International Conference on Social Informatics [27].



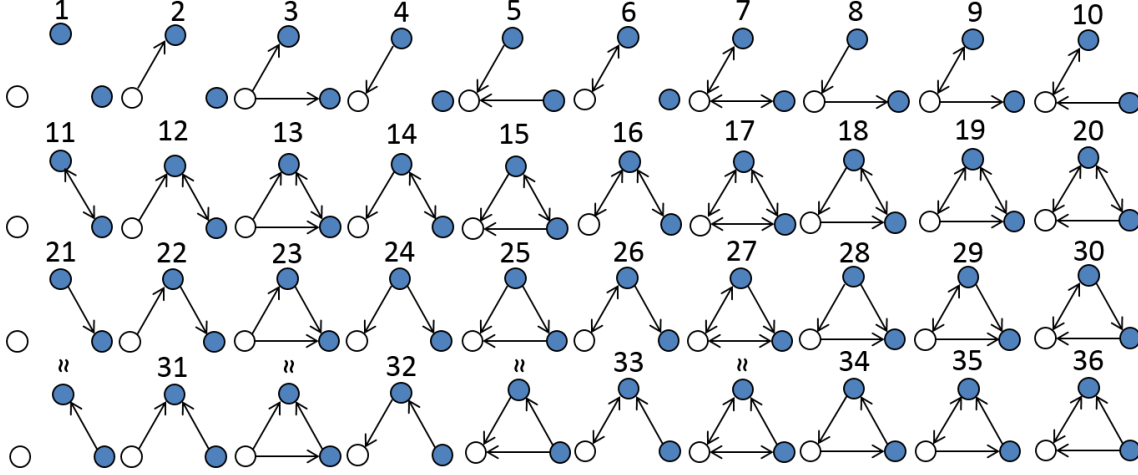


Figure 3: Types of conditional triads

triad [15]. These orientations are the set of all *conditional triads*, which are defined by the structure of the three way relation based on the position of an individual within it. For example, triads 6 and 11 are structurally identical (having two null and one mutual tie). In triad 11, the white user is isolated whereas in triad 6 she is connected to an alter. The entire structure of an ego-network can thus be represented by the number and different types of conditional triads it is composed of. The *conditional triad census* [89, 28] of an ego-network is defined as a 36-element vector whose  $i^{th}$  component represents the proportion of type  $i$  conditional triads it is composed of.

Searching for ego-networks whose conditional triad censuses are similar is expected to lead to a meaningful grouping of users into social roles. This is because each triad configuration represents a sociological factor about how a user interacts with others [17]. For example, triad 32 has a user on the receiving end of a chain of interactions. If these interactions represent the passage of information or rumors, it implies that the alter in the middle of the chain is capable of manipulating what becomes shared with the user and may not be trustworthy. In triad 5, the user receives interactions from two alters but chooses not to reciprocate. Ego-networks largely composed of this triad suggests that the user receives many interactions but, for possibly selfish reasons, seldom chooses to reciprocate. By summarizing how frequently each of these triads appear, a conditional triad censuses succinctly models the strength of the different kinds of social factors that surround the nature of one's interactions with others. These factors, taken together by considering the entire census as a vector, therefore represents the circumstances and reasons why a user participates in a social system.

The number of and kinds of roles that exist in a social system can thus be identified by: (i) computing the conditional triad census of every user; and (ii) clustering users into groups based on the similarity (vector distance) of the conditional triad censuses. This approach is somewhat related to discovering social groups in networks by

Network	$ V $	$ E $	$\bar{d}$	$\bar{C}$	$\alpha_{in}$	$\alpha_{out}$
Facebook	46,952	264,004	37.36	0.085	1.61 ( $p > 0.732$ )	1.68 ( $p > 0.964$ )
Wikipedia	138,592	740,397	10.68	0.038	1.54 ( $p > 0.999$ )	1.83 ( $p > 0.999$ )

Table 1: Dataset summary statistics

searching for ego-networks that participate in similarly shaped  $k$ -cliques [43] or -cores (sub-graphs where all nodes are connected to at least  $k$  others [29]) [48, 76, 56]. However, searching for ego-networks that satisfy these strict requirements will only identify sets of nodes surrounded by a similarly dense network and leave hidden other nodes whose ego-networks are less connected but still have similar connectivity patterns. Such analysis also pays no consideration to the social forces or actions that drive users in cliques or cores to interact with each other, since the types of triads within the groups are ignored. Furthermore, it is difficult to know a priori what kinds of  $k$ -cliques and -cores correspond to relevant social roles in a large-scale social system. In comparison, the proposed approach learns significant structural patterns of ego-networks based on a feature reflecting the types of social forces that bind a user and her connections together. It leads to a classification where users in the same group participate and interact with their contacts under similar social circumstances and forces, which speaks very closely to the notion of a social role.

### 3.2 Dataset description

The methodology is demonstrated by discovering social roles in two popular online social systems, namely Facebook and Wikipedia. These systems were chosen because they each serve a different purpose and provide distinct mechanisms for users to interact with each other. Facebook is used as a platform to informally share personal information, photos, and events with friends and family. Its interaction network is built by placing a directed edge from user  $a$  to  $b$  if  $a$  posts at least one message on the wall (a collection of public messages) of  $b$ . Wikipedia is an online encyclopedia with articles that are written and edited by an open community. Interactions on Wikipedia are defined by the modification of content contributed by another user; a directed edge from  $a$  to  $b$  is added if  $a$  edited the text, reverted a change, or voted on approving an action to an article made by  $b$ . Both the Facebook and Wikipedia networks were constructed from publicly available datasets [87, 64]. These datasets only record the act of an interaction; it does not include any information about the content of or the type of the interaction. Although the Facebook data set is dated (interactions were recorded in 2009), privacy improvements made to the Facebook API since make it all but impossible to capture such interactions at scale today.

Table 1 presents summary statistics for these interaction networks, illustrating how they vary in size, shape, and

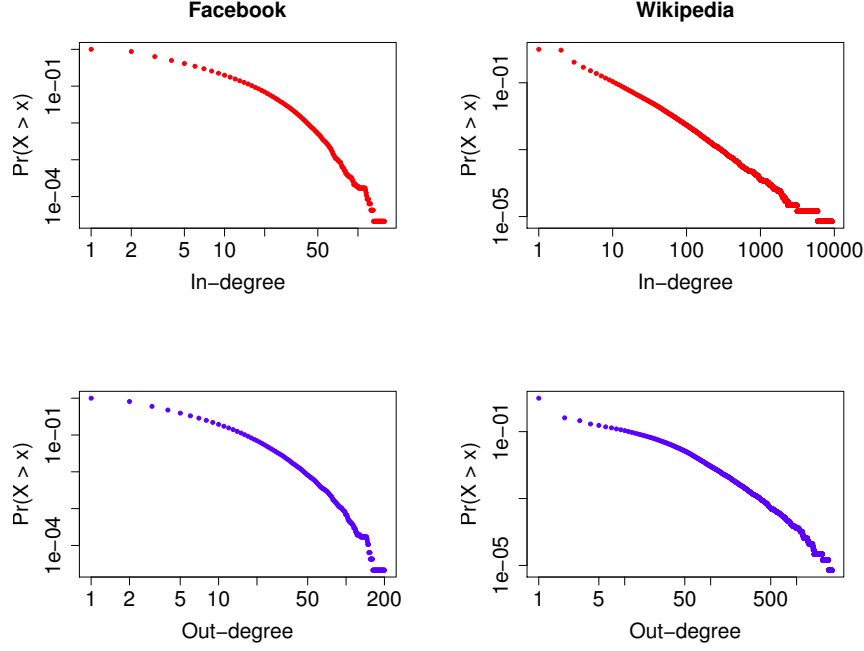


Figure 4: In- (top) and out-degree (bottom) distributions

user behaviors. The relatively small size (46,952 users, 264,004 pairwise interactions) of the Facebook network is due to the fact that it only represents users within a single regional network (the Facebook social graph was divided by user regions in its earliest form). The data also only represents user's whose accounts were shared publicly, which was the default Facebook setting during the data collection period [87]. Despite its size and limit to a single regional network, previous work showed that all regional Facebook networks exhibited a similar structure (average path lengths and diameter) and shape (clustering coefficients and assortativity) [96]. More recent studies further confirm that the structure and shape of these regional networks are very similar to the structure of the modern global Facebook network [97, 86]; therefore this data set is expected to contain similar interaction patterns as seen in the global Facebook network. The Wikipedia network is almost three times the size of Facebook, with 138,592 users and 740,397 distinct pairwise interactions, but its clustering coefficient  $\bar{C}$  is approximately 55% smaller. These measurements suggest that Facebook users have a greater tendency to surround themselves within denser ego-networks compared to Wikipedia users. The lower clustering coefficient of Wikipedia could be explained by users who generally limit themselves to modifying articles written by a specific group (perhaps representing a specific topic).

The in- and out-degree distributions of each network is presented in Figure 4, which exhibit power-tailed shapes. The existence of power-law behavior is tested by a maximum likelihood approach [21] and the resulting power-

law exponents  $\alpha_{in,out}$  are given in Table 1. The estimates of the power-law exponent are very reliable ( $p > 0.95$ ; note that the test considers the hypothesis  $H_0$ : the empirical data follows a power-tailed distribution) except for the in-degree distribution of Facebook, which may be because its range only covers two orders of magnitude. A larger power-law exponent indicates that the distribution drops to zero faster in its right-tail [61], hence the frequency with which users interact with others on Wikipedia exhibits a smaller amount of variation compared to Facebook. In other words, it is less likely to find a user who interacts with an unexpectedly high number of others on Wikipedia compared to Facebook, and less likely to find a user receiving many interactions from others on Facebook compared to Wikipedia.

### 3.3 Network sampling

Computing the conditional triad census of every ego-network requires an examination of  $O(|V|^3)$  triples of users in an interaction network. This computational cost may be an insurmountable burden to compute conditional triad censuses in larger interaction networks where the number of nodes are in the millions [86]. Furthermore, existing algorithms that can compute censuses in  $O(|V|^2)$  [68] or  $O(|E|)$  [5] only considers users' *unconditional* triad censuses. An unconditional triad census is a 16-element vector holding the proportion of all triads without regard to the position of the user in her ego-network, making them incompatible with the proposed approach. However, since the components of a conditional triad census are the *proportions* of triad types in an ego-network, the conditional censuses within a carefully selected *sample* of the original network should be representative of the conditional censuses in the original network. A sample of a network  $G$  is a new network  $G_s = (V_s, E_s)$  where  $V_s \subset V$ ,  $E_s \subset E$ , and  $|V_s| = \phi|V|$  with  $0 < \phi < 1$ .

A sampling method must ensure that the two critical local structural properties of ego-networks, namely the degree distribution and local clustering coefficient distribution are preserved [46, 31]. For example, ego-networks with high degree will naturally tend to have triads with relations among multiple alters, and lower (higher) cluster coefficients indicate a greater proportion of open (closed) triads. However, naïve methods for network sampling do a poor job of preserving these local features. A number of advanced sampling methods have been proposed, but each one can only preserve different types of structural features of the full network [1]. Therefore, four widely used graph sampling techniques for choosing  $V_s$  and  $E_s$  were compared by their ability to preserve the degree distribution of the users' ego-network and their clustering coefficients. The techniques and their process are:

1. **Vertex Sampling (VS)**: Let  $V_s$  be a random sample of  $\phi|V|$  vertices from  $V$  and define  $E_s$  to be the set of all edges among the vertices in  $V_s$  from  $G$ .
2. **Edge Sampling (ES)**: Randomly choose an edge  $e = (v_1, v_2)$  from  $E$ , add it to  $E_s$ , and add  $v_1$  and  $v_2$  to  $V_s$

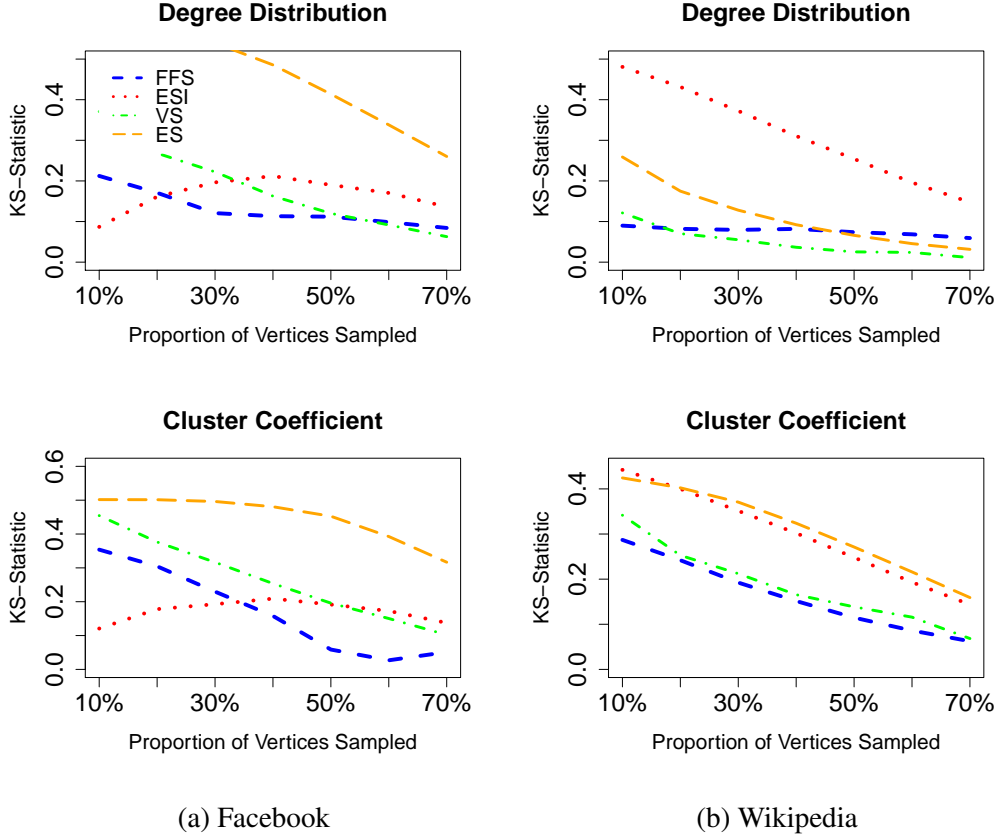


Figure 5: Comparison of graph sampling methods

if they have not yet been added. Continue to choose edges from  $E$  until  $|V_s| = \phi|V|$ .

3. **Forest Fire Sampling (FFS)** [59]: Choose a random vertex  $v$  from  $V$ , randomly select  $p/(1-p)$  of its outgoing edges, and add these edges to  $E_s$ . Place every vertex incident to those added to  $E_s$  into a set  $V_*$  of ‘burned vertices’ and update  $V_s$  by  $V_s = V_s \cup V_*$ . Randomly choose a burned vertex from  $V_*$ , and recursively repeat this process until  $|V_s| = \phi|V|$ . The parameter assignment  $p = 0.7$  is used based on the recommendation of the method’s authors [59].
4. **ES-i (ESI)** [1]: Randomly choose an edge  $e = (v_1, v_2)$  from  $E$  and add  $v_1$  and  $v_2$  to  $V_s$  if they have not yet been added (note that  $e$  is not added to  $E_s$ ). Continue sampling until  $|V_s| = \phi|V|$ . Finally, define  $E_s$  to be the set of all edges among the vertices in  $V_s$  from  $G$ .

The Kolmogorov-Smirnov distance metric  $D$  was used to compare how closely the degree and clustering coefficient distributions of samples  $F_s$  taken with each method follow the distribution of the original network  $F$ . It is

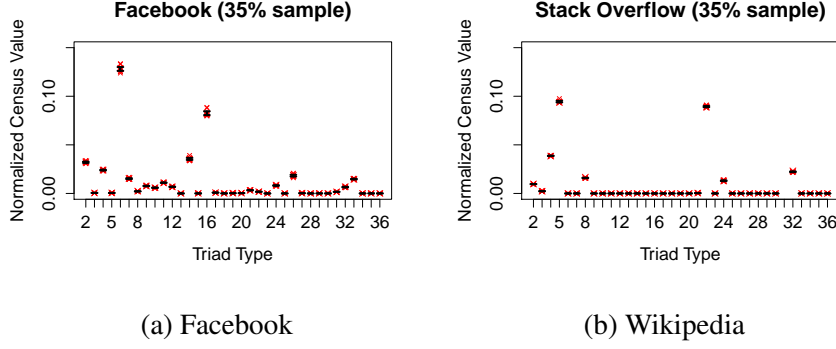


Figure 6: Triad Role Census Sample Values with 95% Confidence Intervals

defined by the largest difference of a point taken from original distribution  $F$  to the distribution of the sample  $F_s$ :

$$D = \sup_x |F_s(x) - F(x)|$$

Figure 5 compares the average  $D$  over 100 samples taken by each method for different values of  $\phi$ . For the Facebook network, FFS does the best job ( $D < 0.2$ ) at preserving both degree and clustering coefficient distributions for modest sample sizes ( $\phi \geq 0.33$ ). FFS samples of the Wikipedia network best preserves the clustering coefficient distribution for any sample size and  $0.2 < \phi < 0.3$  FFS and VS are similarly faithful to the original network’s degree distribution. Ultimately, FFS sampling is found to be able to preserve the local structure of both networks even for small sample sizes.

A value of  $\phi$  that provided a reasonable trade-off between computational speed and sample consistency was searched for. Figure 6 plots the average value of each component of conditional triad censuses taken from  $n = 20$  independently generated FFS samples of each network for  $\phi = 0.35$  (triad 1 is excluded because of its disproportionately high frequency) and the 95% confidence interval of the proportions. The proportion of triad types across the samples are similar and feature small confidence intervals. Since the computation cost of computing triad censuses at this sampling level is very reasonable (less than 30 minutes in a parallel computation over three cores of an Intel i5 processor), the setting  $\phi = 0.35$  is used for role analysis.

### 3.4 Census clustering

$k$ -means clustering, a common and flexible algorithm for discovering latent groups in data [45, 100, 37], is used to separate users into roles.  $k$ -means clustering defines  $k$  centroid positions in the vector space and assigns each conditional triad census (and hence user) to a cluster based on the centroid it is most similar to. Since the components of the censuses take a value between 0 and 1, this similarity is defined as the  $\ell_2$ -norm of their

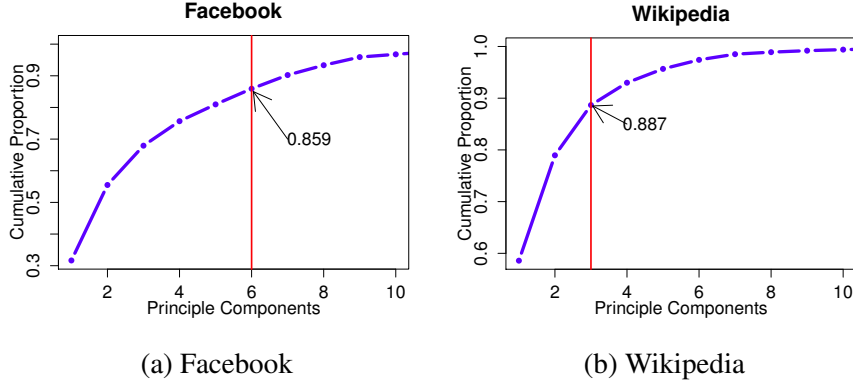


Figure 7: Scree plots

difference vector. After the assignment of conditional triad censuses to clusters, the position of the centroid of each cluster is updated. Censuses are then reassigned to their closest centroid, and the process repeats until there are no changes to any cluster assignments.

### 3.4.1 Dimensionality reduction

Figure 6 indicates that many components of the conditional triad censuses are close to or equal to 0. A dimensionality reduction technique, namely principle component analysis (PCA) [47], is therefore applied to the conditional triad censuses. PCA identifies a projection of the data into a lower dimensional subspace that preserves as much variation within the original space as possible. Figure 7 plots the proportion of variation within the original dataset that is retained when we use PCA to reduce the data into smaller numbers of principle components. The smallest dimensional space that still preserved a large proportion of the variation in the data ( $> 85\%$ ) was chosen, as indicated by the red line in Figure 7. The figure suggests that PCA finds a significantly lower dimensional space for clustering the conditional triad census of every network, from 36 dimensions to just 6 and 3 for the Facebook and Wikipedia interaction networks respectively.

### 3.4.2 Clustering evaluation

$k$ -means clustering requires the number of clusters  $k$  to divide the data into to be chosen beforehand, forcing an analyst to assert the specific number of social roles that may exist in the system. Instead, the silhouette coefficient metric [83]  $SC_{\hat{C}^k}$  is used to quantitatively evaluate the quality of clusters for different values of  $k$ , so that the  $k$  yielding the ‘best’ clustering is chosen. It is defined as follows: consider a division of censuses into  $k$  clusters  $\hat{C}^k = \{C_1, C_2, \dots, C_k\}$ . Let  $\alpha(\mathbf{x}) = d(\mathbf{x}, C_i^*), x \in C_i$  be the distance from the vector  $\mathbf{x}$  to the centroid  $C_i^*$  of its

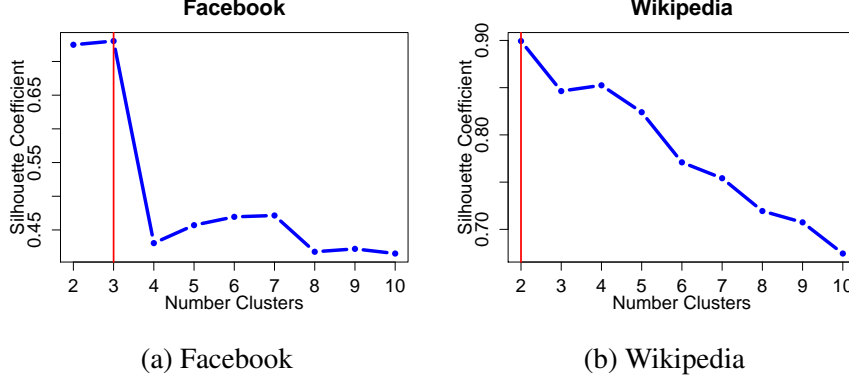


Figure 8: Silhouette coefficients

assigned cluster  $C_i$  (measuring intra-cluster distance) and  $\beta(\mathbf{x}) = \min_{C_j \in \hat{C}^k, C_j \neq C_i} d(\mathbf{x}, C_j^*)$  be the distance from  $\mathbf{x}$  to the centroid of the nearest cluster  $C_j$   $\mathbf{x}$  is not assigned to (measuring inter-cluster distance). The silhouette of  $\mathbf{x}$  is defined as:

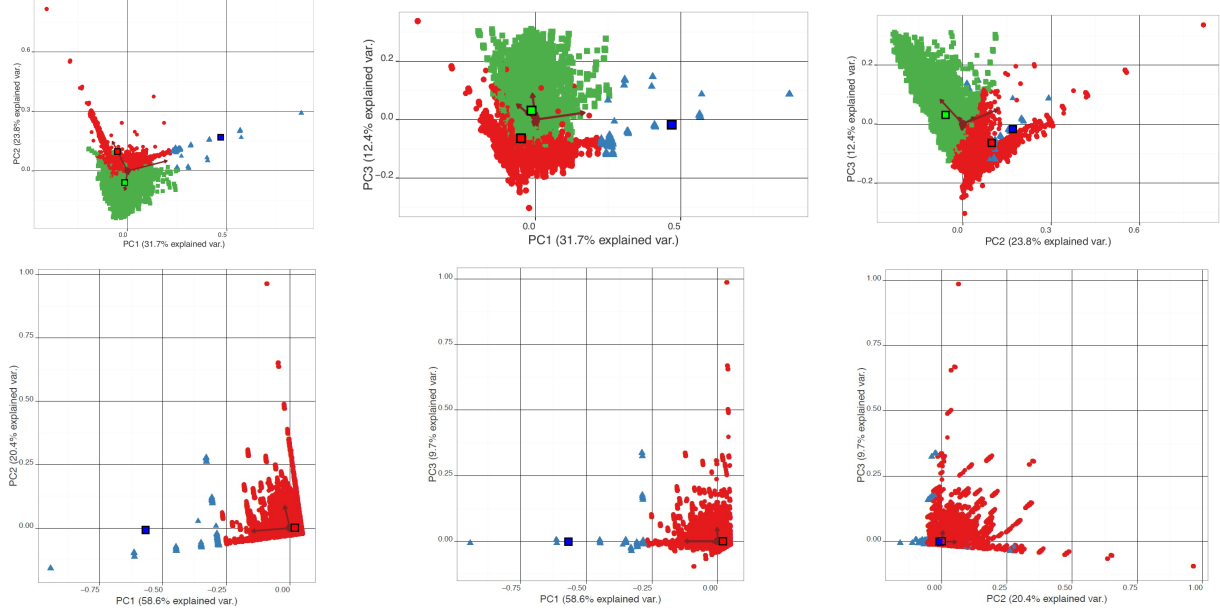
$$\phi(\mathbf{x}) = \frac{\beta(\mathbf{x}) - \alpha(\mathbf{x})}{\max(\beta(\mathbf{x}), \alpha(\mathbf{x}))}$$

Note that  $\phi(\mathbf{x})$  approaches 1 as the separation between the cluster  $\mathbf{x}$  is assigned to and the nearest other cluster increases. The average silhouette of every clustered vector defines the silhouette coefficient of a clustering  $\hat{C}^k$ :

$$SC_{\hat{C}^k} = \frac{\sum_{\mathbf{x} \in \mathbf{X}} \phi(\mathbf{x})}{|\mathbf{X}|}$$

where  $\mathbf{X}$  is the set of all data vectors. Previous studies indicate that values of  $SC_{\hat{C}^k}$  greater than 0.7 means the algorithm achieved superior separation, and values between 0.5 and 0.7 indicate a reasonable separation [83]. For a given value of  $k$ , we ran 50  $k$ -means clusterings over the PCA-reduced conditional triad censuses using different random initializations of the centroid positions. Figure 8 plots the average  $SC_{\hat{C}^k}$  of these trials for  $2 \leq k \leq 9$ . It reveals excellent clustering solutions at  $k = 3$  and  $k = 2$  clusters for the Facebook and Wikipedia censuses, with silhouette coefficients of 0.73 and 0.90, respectively. A qualitative validation of the adequacy of a clustering solution is also given in Figure 9. Here, the conditional triad censuses in a space defined by the first three principle components are assigned a marking and color corresponding to their cluster assignment. The Facebook clustering solution, given in the top panels of the figure, discovers a role (the red cluster of circle points) that exhibits large variation along two principle components. In contrast, a second role (the green cluster of square points) varies strongly along the third component. The smallest cluster (blue cluster of triangle points) only varies along the first component. Since the clusters exhibit little variation along different directions, different subsets of conditional triads must appear in similar proportions within the censuses of the same group. The Wikipedia





(a) Principle Comp. 1 and 2

(b) Principle Comp. 1 and 3

(c) Principle Comp. 2 and 3

Figure 9: Clusters along the first three principle components for Facebook (top) and Wikipedia (bottom)

clustering solution, given in the bottom panels of Figure 9, also finds that the two clusters vary along the direction of different principle components: the red cluster of circle points vary along the second and third components, while the blue cluster of triangle points mainly varies along the first component.

#### 4 Triad-based Role Analysis

In this section, the kinds of social roles that emerge from our clustering analysis is analyzed. For this purpose, the average centroid positions  $C_i^*$  over a clustering result was identified and the user  $u_i^*$  whose conditional triad census is located closest to  $C_i^*$  was found.  $u_i^*$  is defined as the “central user” of role  $i$  whose ego-network is the “central structure” of the role. Due to its position in the cluster, this “central structure” represents the way a prototypical user having this role embeds herself within the social system. In other words, the ego-network structure of users in role  $i$  are most similar to  $C_i^*$  compared to any other central structure on the network. Each central structure is given a social role label based on a subjective interpretation of the user’s position within it. The label captures the way users of a role interact with others in the system, and how the structure representing a role affects the kinds of interactions that are possible. The role labels may not be applicable to all social systems, although it is feasible that systems created under a similar context (e.g. social sharing sites) exhibit similar central role structures and labels. The central role structures discovered in the Facebook and Wikipedia networks, and

Role label	Structure	Proportion of users
Social Group Manager	Figure 10(a)	56.6%
Exclusive Group Participant	Figure 10(b)	28.4%
Information Absorber	Figure 10(c)	15.0%

Table 2: Facebook roles

support for the emergence of these roles in the literature, are presented next.

#### 4.1 Facebook

Figure 10 presents the central role structures of the three social roles found on Facebook. A label representing each role structure and the proportion of users falling under each are presented in Table 2. In these figures, the red node (with a red arrow pointing to it) corresponds to the central user and the blue nodes are the members of her ego-network. The structure in Figure 10(a) represents a social role the majority of all Facebook users (56.6%) fall into where a user is centrally embedded between many disconnected groups of others. She lies in a position critical for maintaining connectivity between communities, and hence, lies in the brokerage position of many open triads. These many open triads give users in this role many opportunities to control if and how information exchanges from one group to another. However, given the fact that Facebook is used as a platform for social sharing, such users may never decide to share information between communities when they represent different social circles. For example, one can envision the user in Figure 10(a) to be sitting between groups that may correspond to colleagues at work, relatives, personal friends, and work colleagues. A user may never want personal information shared among relatives to be revealed by work colleagues, and may want conversations, rumors, and other information shared among friends to never be exposed to family members and work colleagues. That a majority of Facebook users fall into a social role that brokers among many disconnected circles is not surprising; many past research studies have shown that most Facebook users face identity management and multiple presentation issues while interacting on the site [26, 81, 4, 57]. Identification of these “social group managers” is thus a way of finding the bridges or weak ties [14] in the network based on structural patterns rooted in social theory.

28.4% of Facebook users fall into the role represented by the central structure of Figure 10(b). This structure represents a user that has surrounded herself around a web of interactions running between her first-degree connections. This small percentage of users only participates in a single, tight-knit community of others rather than managing many disconnected groups. Such a role may represent users who only choose to ‘friend’ and interact with a collection of others that share many mutual connections, and does not need to manage multiple discon-

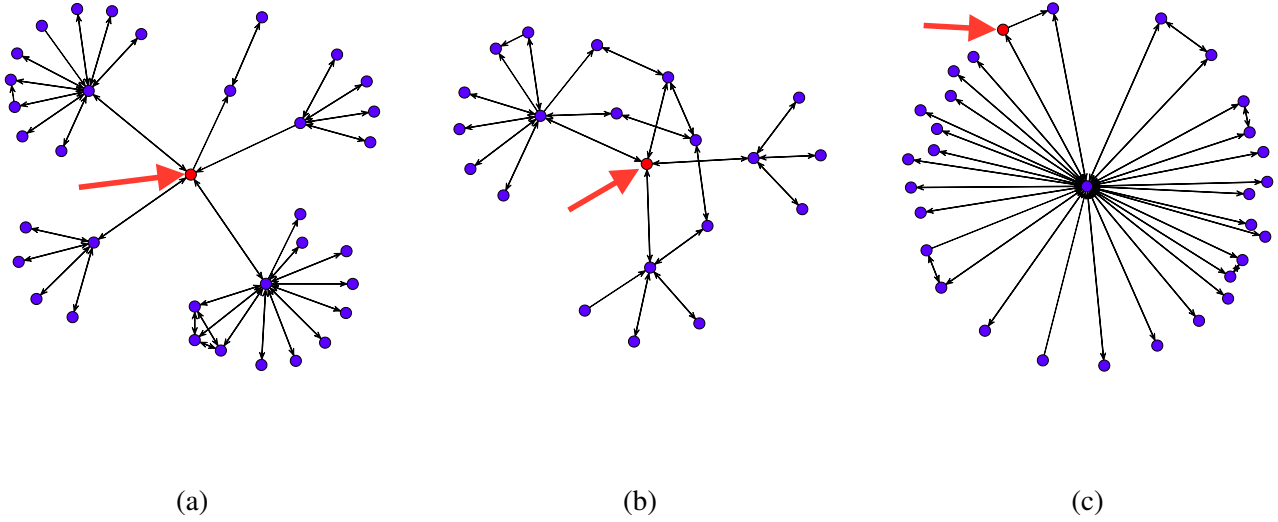


Figure 10: Central role structures: Facebook

nected social circles. Such patterns are known to be more prominent in the ego-networks of Facebook users who are more willing to share information with many others, or does not feel a need to consider identity management on the social network [42, 65, 74, 12]. Such “exclusive group participants” may therefore promote the use of Facebook as a genuine social sharing platform, and be instrumental in the development of dense interaction clusters in the structure of the network.

Figure 10(c) corresponds to the 15% of users who are positioned at the periphery of a single alter that interacts with many others. Since the structure corresponds to an average or typical ego-network structure for users in this role, it signifies a group of users who are passive and seldom share information with others. When they do share, it tends to be with those who the user has a mutual association with. Furthermore, these users tend to receive information from alters that share prolifically. The phenomenon of over-active or extraordinarily well connected users on online social systems is well-studied [67, 54, 55], but it is interesting to discover that the users connected to them to also play an important role in the online system. These users ‘absorb’ the information of the over-active others, since they only forward such information to those already connected to the over-active source. In fact, a modern use of the Facebook platform is to “absorb information” from friends and news organizations rather than to share social information, as reflected by this social role [3, 80, 88].

Role label	Structure	Proportion of users
Interdisciplinary Contributor	Figure 11(a)	89.7%
Technical Editor	Figure 11(b)	10.3%

Table 3: Wikipedia roles

## 4.2 Wikipedia

The density of the central structures of the Wikipedia social roles shown in Figure 11 is a result of the many different ways interactions are defined, which includes content editing, reverting a change, or voting on a pending action by another user. The triad-based analysis revealed two types of roles in Wikipedia. The first role is taken on by the majority of all users (89.7%) and has the central structure shown in Figure 11(a). The structure shows a user whose work is being changed by active alters that make changes to articles from many other authors as well. It is interesting that these the active alters seldom edit content added by a common individual (e.g. have few mutual connections), even though they are prolific editors. Such a pattern may emerge when these alters have different expertise and concentrate on editing contributions that fall within their specific domain. The existence of these ‘hubs’ of editing activity is not a surprising finding, as past work has confirmed that most editors on Wikipedia do exhibit domain-specific expertise and limit their edits articles in their domain [92]. Users falling under this structure must therefore be contributing to interdisciplinary articles, which most Wikipedia articles are classified as [66]. Such “interdisciplinary contributors” represent the vast majority of users (89.7%) and is thus the primary role that adds information to Wikipedia.

The remaining 10.3% of users fall under the role whose central structure is given in Figure 11(b). Two alters that take the form of a hub (a domain-specific expert) can be seen, but the overlap between them is larger and denser in comparison to Figure 11(a). The central user is positioned within this overlap. Users in this role therefore edit the contributions of many, and find their contributions edited by many others as well. A plausible explanation for finding a dense core between the positions of domain experts is that they perform ‘general’ edits that reflect the language, grammar, spelling, hyperlinking, and structure of articles. Changes made by these “technical editors” may be further refined by a large number of other editors to further refine the technical discussion or the presentation and language of an article. This explanation is compatible with past observations of users that concentrate on edits related to the language and format of an article [92].

In summary, the analysis demonstrates the use of conditional triad censuses to extract social roles from different

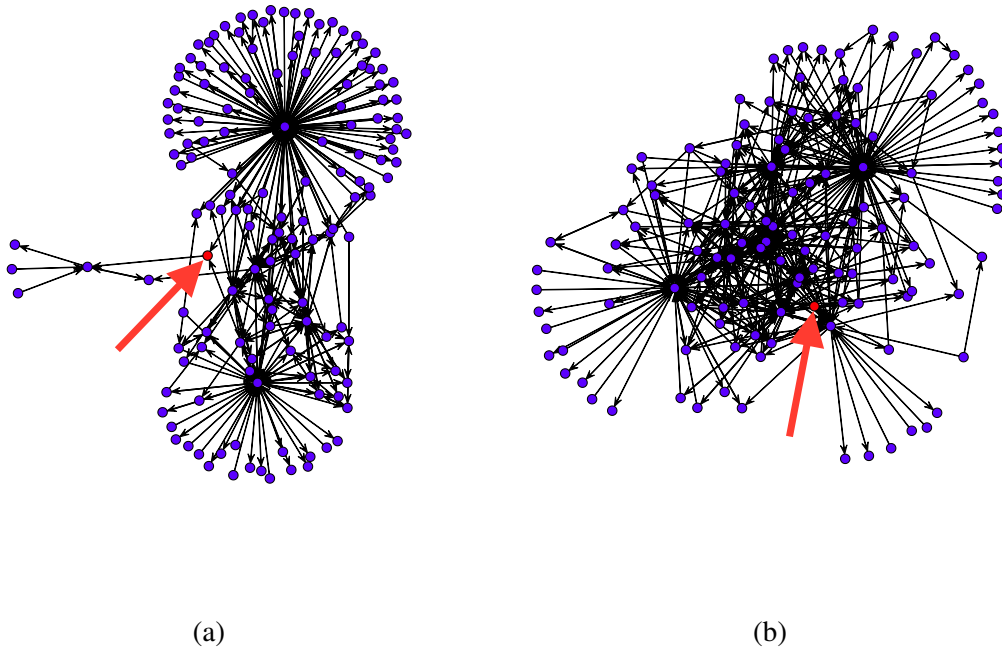


Figure 11: Central role structures: Wikipedia

types of online social systems. It naturally discovered social roles on Facebook corresponding to users who maintain connectivity across many disconnected social groups (“social group managers”), who participate in well-connected groups (“exclusive group participants”) that generate many social interactions, and passive users that serve as an outlet for over-active others to share information with (“information absorbers”) and may use Facebook as a platform to receive news. The roles discovered on Wikipedia focuses on the nature of the user’s contribution to the content of the online encyclopedia. A majority of users (“interdisciplinary contributors”) are devoted to articles that attract the attention of editors focusing on different subsets of articles, which may correspond to the actions of a domain-specific expert. The attraction of many experts suggests that the article the central user focuses on is interdisciplinary in nature. A minority of users (“technical editors”) edit many articles at once, and have their articles edited by many others as well. These users may thus be domain-specific experts or could be users that apply general language and formatting changes to many articles on the site.

### 4.3 Applying social role analysis

Triad-based social role analysis offers not only insights into the nature of user behaviors on social systems, but also a practical tool for exploring social theories. For example, consider a researcher wishing to study whether or

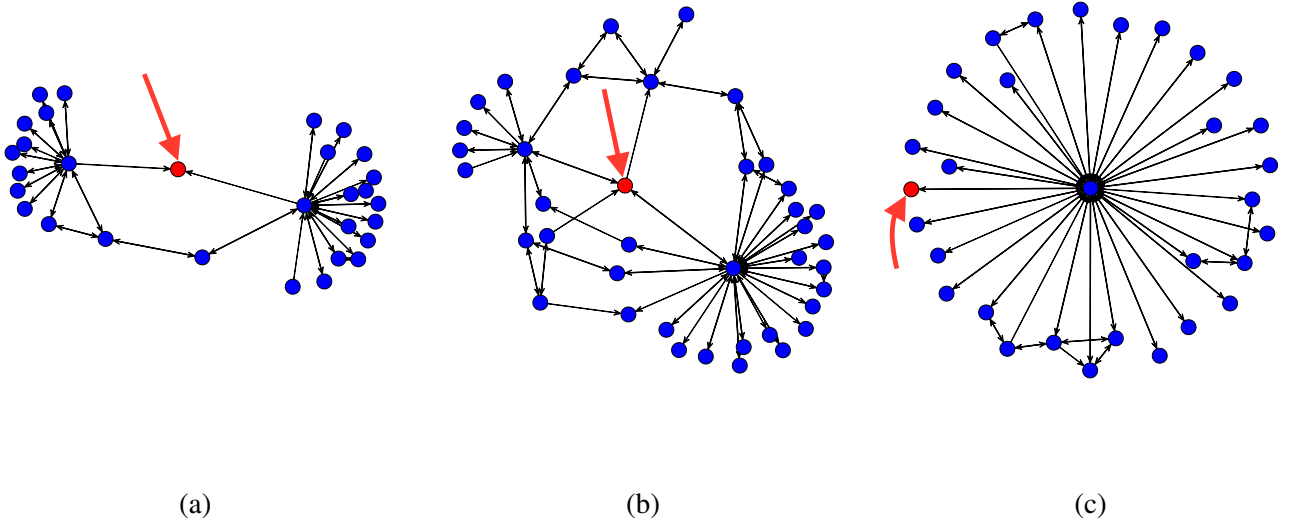


Figure 12: Central role structures: UC Irvine online network

not the reasons and ways users interact with each other on Facebook is due to its inclusive, public nature. This research question may be explored by comparing the social roles that emerge on Facebook with a different online social network that is not inclusive and public, but exclusive and private. Differences in the number, shape and proportion of users falling into social roles across the two systems may give evidence of a relationship between the public or private nature of a social system and why people participate in it. To illustrate this, a data set of interactions recorded from a private online social network for students at the University of California Irvine (UC Irvine) is considered [72]<sup>2</sup>. The six-month long data set consists of 1,899 ties between users, with a directed tie from user  $A$  to  $B$  established when  $A$  sends at least one message to  $B$ . Triad-based social role analysis on the UC Irvine network revealed the best clustering solution at  $k = 3$  roles ( $\hat{C}_k = 0.713$ ). Figure 12 visualizes the central structure of the resulting role clusters, which exhibit very similar features to the central structures of the social roles on Facebook. For example, Figures 10(a) and 12(a) both have a user situated between two groups of others, Figures 10(b) and 12(b) find the user in the center of a well connected community, and Figures 10(c) and 12(c) shows the user sitting at the periphery of a highly active alter. An analyst may therefore consider the two networks to exhibit the same social roles, and hence, conclude users utilize the network for similar reasons and in similar ways. Given the fact that Facebook and the UC Irvine social networks were created to facilitate social

<sup>2</sup>It should be emphasized that a complete study of this research question requires a comprehensive analysis of user behaviors, and extensive comparisons between many different social network datasets. The illustration that follows is limited, and is only meant to demonstrate how social role analysis can be used as a useful research tool.

Role Label	Proportion of Users	
	UC Irvine	Facebook
Social Group Manager	3.06%	56.6%
Exclusive Group Participant	92.9%	28.4%
Information Absorber	4.04%	15.0%

Table 4: Comparing the proportion of social roles on Facebook and UC Irvine networks

communication and connection, it is not surprising to find similar roles and central structures emerging.

Comparison of the shape and the proportion of users falling into the central role structures, however, reveal significant differences between the private UC Irvine and public Facebook online social networks. For example, the central role structure of social group managers in the UC Irvine network finds the ego to be situated between a smaller number of groups compared to Facebook, and has an additional alter managing the same set of social groups. These differences may arise because the separate groups an individual participates in within a private social network that is smaller in scope and encompasses fewer types of people may be less than a public social network that can include family, social, and work contacts. Furthermore, Table 4 shows the proportion of users falling into the social roles of the two networks to be very different. The majority of users in the UC Irvine network are exclusive group participants, that is, they are found to be embedded within a tight social group and do not need to manage a membership in many separate ones. In fact, only 3.06% of UC Irvine users act as a social group manager, compared to the 56.6% of Facebook users that take on this role. This difference may be rooted in the fact that its users are all students of UC Irvine, and hence, may exhibit homopholic tendencies through common class, standing, housing, major, college, and club affiliations. The many ways by which users could exhibit homophily on the UC Irvine network may also explain why the social group manager central structure has an alter managing the same set of groups as the ego; both could be managing groups of colleagues from the same class and club. The public nature of Facebook, however, may be reducing the level of homophily among a user’s connections. An analyst may point to these findings as key differences between public and private online social networks, and as a rationale to explore new hypotheses involving a comparison of homopholic tendencies within them.

## 5 Further Opportunities for Large Scale Social Role Analysis

Based on the related work discussed in Section 2 and on a reflection of the proposed triad-based method, this section summarizes additional challenges and opportunities that exist in social role analysis for large scale social systems. Opportunities along two important directions are considered: (i) finding meaningful features for role

extraction; and (ii) understanding the relationship between functional and social roles.

### 5.1 Linking representation with social theory

As discussed in Section 2.4, many data-driven analyses select a large collection of structural, user, and relationship attributes, and use them all to discover the social roles within a system. However, this may be a dangerous practice because the resulting roles are defined to be according to the ‘similarity’ of a complex mixture of many variables. Furthermore, many quantitative structural, user, and relationship features do not necessarily have a close correspondence to a sociological theory that is related to the concept of a user role. For example, structural features such as the clustering coefficient or betweenness centrality of a user within her ego-network can quantify how clustered its structure is, but does not identify the telling patterns of the interactions within it. Analysis that use a large number of features thus lead to a separation of users into roles that must be defined very broadly, or where ego-network structures within roles may be discordant and have few interpretable structural regularities. Some methods using a large collection of features also apply post-processing steps to the resulting groups [18, 102], which may further distort any interpretation of the extracted roles.

This article takes a step toward the exclusive use of features that carry a specific social interpretation. However, it may be the case that additional features associated with social theories may improve the fidelity of the method’s results, or that a different unsupervised learning algorithm should be used. For example, Field *et al.* note the importance of preserving not only interactions, but also affiliation information between users in a social system to define their position [34]. Such a concept may be operationalized in a richer dataset containing affiliation information, by incorporating similarity measures of the rows of a  $g \times n$  binary incidence matrix whose  $i^{th}$  row and  $j^{th}$  column is 1 if user  $j$  is affiliated with group  $i$ . Another related concept is the importance of social influence to the way it impacts a user’s social role [35]. Fortunately, there have been many measures proposed for quantifying influence that may be integrated into the social role mining process [73, 36, 19, 60, 20, 51]. It is these kinds of factors, instead of conveniently chosen structural and user features, that should be considered when grouping users into social roles.

### 5.2 Linking functional and social roles

In an offline setting, people can interact, converse, and exchange ideas with each other in virtually innumerable ways. However, most large scale social datasets come from online systems that only offer a limited number of well-defined ways for people to interact with one another. It may be intuitive to think that these modes of interactions, which reflect the *functional* ways users participate on the social system, are associated or have an



effect on the social roles they go on to exhibit. For example, the roles identified over Wikipedia in this article were more closely related to the types of interactions allowed by the service (as an expert editing content or a generalist editing language form). The functionality provided by Facebook may also have helped users fall into specific social roles; for example, users only participating in a cooperative group of others may leverage the ability to choose what friendships to accept on Facebook, so that the group they are embedded in is cohesive. The idea that users can only interact with others in a limited number of ways is a unique property of online social systems compared to offline ones. Thus, advanced features used to discover social roles may also need to be associated with the different functionalities of an online social service, with values that reflect what functions and how frequently they are used. Such features that are found to be ‘significant’ across classes of users falling under the same social role may signal an association between the functionality of a social system and its social roles.

## 6 Concluding Remarks

This article presented a methodology to discover social roles in large scale social systems. The data-driven approach, rooted in the representation of ego-networks as a conditional triad census and implemented with a simple unsupervised learner was applied to two different online social systems. Structural analysis of the ego-networks falling closest to the center of clusters of users with similar conditional triad censuses suggested the presence of users on Facebook that exclusively manage disconnect social circles or participate in a highly collaborative singular one. It also found how content posted on Wikipedia may attract either the attention of a number of domain experts, or of multiple generalist editors. The data-driven approach was motivated by a comparative analysis of the existing equivalence based, implied, and data-driven role discovery methods that had been proposed. It concluded by suggesting the integration of social theories to derive features for role mining, and approaches to link together the notion of what a user can do on a social system with her social role on it. Future work should explore these opportunities, and may also consider unsupervised learners that allow users to fall into multiple role assignments. It is hoped that this important topic will continue to gain more attention in the computational social network analysis and mining community.

## References

- [1] N. K. Ahmed, J. Neville, and R. Kompella. Network sampling: from static to streaming graphs. *arXiv preprint arXiv:1211.3412*, 2012.
- [2] A. S. Alderson and J. Beckfield. Power and position in the world city system<sup>1</sup>. *American Journal of sociology*, 109(4):811–851, 2004.

- [3] B. Baresch, L. Knight, D. Harp, and C. Yaschur. Friends who choose your news: An analysis of content links on facebook. In *The Official Research Journal of International Symposium on Online Journalism*, volume 1, pages 1–24, 2011.
- [4] L. Barkhuus and J. Tashiro. Student socialization in the age of facebook. In *Proc. of the SIGCHI Conference on Human Factors in Computing Systems*, pages 133–142. ACM, 2010.
- [5] V. Batagelj and A. Mrvar. A subquadratic triad census algorithm for large sparse networks with maximum degree. *Social Networks*, 23(3):237–243, 2001.
- [6] B. Batjargal. Network triads: transitivity, referral and venture capital decisions in china and russia. *Journal of International Business Studies*, 38(6):998–1012, 2007.
- [7] A. Baum, A. Shapiro, D. Murray, and M. V. Wideman. Interpersonal Mediation of Perceived Crowding and Control in Residential Dyads and Triads. *Journal of Applied Social Psychology*, 9(6):491–504, 1979.
- [8] S. P. Borgatti and M. G. Everett. The class of all regular equivalences: Algebraic structure and computation. *Social Networks*, 11(1):65–88, 1989.
- [9] S. P. Borgatti and M. G. Everett. Notions of position in social network analysis. *Sociological methodology*, 22(1):1–35, 1992.
- [10] S. P. Borgatti and M. G. Everett. Regular blockmodels of multiway, multimode matrices. *Social Networks*, 14(1):91–120, 1992.
- [11] S. P. Borgatti and M. G. Everett. Two algorithms for computing regular equivalence. *Social Networks*, 15(4):361–376, 1993.
- [12] T. E. Bosch. Using online social networking for teaching and learning: Facebook use at the university of cape town. *Communicatio: South African Journal for Communication Theory and Research*, 35(2):185–200, 2009.
- [13] D. Brass, K. Butterfield, and B. Skaggs. Relationships and Unethical Behavior: A Social Network Perspective. *The Academy of Management Review*, 23(1):14–31, 1998.
- [14] M. Burke and R. Kraut. Using facebook after losing a job: Differential benefits of strong and weak ties. In *Proc. of the ACM Conference on Computer Supported Cooperative Work*, pages 1419–1430. ACM, 2013.
- [15] R. S. Burt. Detecting role equivalence. *Social Networks*, 12(1):83–97, 1990.

- [16] E. Cambria, D. Rajagopal, D. Olsher, and D. Das. Big social data analysis. *Big data computing*, pages 401–414, 2013.
- [17] T. Caplow. *Two against one: Coalitions in triads*. Prentice-Hall Englewood Cliffs, NJ, 1968.
- [18] J. Chan, C. Hayes, and E. M. Daly. Decomposing discussion forums and boards using user roles. In *Intl. Conference on Weblogs and Social Media*, volume 10, pages 215–218, 2010.
- [19] W. Chen, Y. Wang, and S. Yang. Efficient influence maximization in social networks. In *Proc. of 15th ACM SIGKDD Intl. Conference on Knowledge discovery and data mining*, pages 199–208. ACM, 2009.
- [20] W. Chen, Y. Yuan, and L. Zhang. Scalable influence maximization in social networks under the linear threshold model. In *Proc. of 10th IEEE Intl. Conference on Data Mining*, 2010.
- [21] A. Clauset, C. R. Shalizi, and M. Newman. Power-Law Distributions in Empirical Data. Technical report, arXiv:0706.1062v2 [physics.data-an], 2009.
- [22] K. S. Cook and R. M. Emerson. Power, equity and commitment in exchange networks. *American Sociological Review*, pages 721–739, 1978.
- [23] K. S. Cook, R. M. Emerson, M. R. Gillmore, and T. Yamagishi. The distribution of power in exchange networks: Theory and experimental results. *American journal of sociology*, pages 275–305, 1983.
- [24] J. Davis and S. Leinhardt. The Structure of Positive Interpersonal Relations in Small Groups. *Sociological Theories in Progress*, 2:218–251, 197.
- [25] P. DiMaggio. Structural analysis of organizational fields: A blockmodel approach. *Research in organizational behavior*, 8:335–370, 1986.
- [26] J. M. DiMicco and D. R. Millen. Identity management: multiple presentations of self in facebook. In *Proc. of ACM Intl. Conference on Supporting Group Work*, pages 383–386. ACM, 2007.
- [27] D. Doran. Triad-based Role Discovery for Large Social Systems. In *Proc. of Intl. Conference on Social Informatics Workshops, LNCS 8852*, pages 130–143, 2014.
- [28] D. Doran, H. Alhazmi, and S. Gokhale. Triads, Transitivity, and Social Effects in User Interactions on Facebook. In *Proc. of IEEE Intl. Conference on Computational Aspects of Social Networks*, pages 68–73, 2013.

- [29] S. N. Dorogovtsev, A. V. Goltsev, and J. F. F. Mendes. K-core organization of complex networks. *Physical review letters*, 96(4):040601, 2006.
- [30] B. H. Erickson. The relational basis of attitudes. *Social structures: A network approach*, 99:121, 1988.
- [31] G. Fagiolo. Clustering in complex directed networks. *Physical Review E*, 76(2):026107, 2007.
- [32] T.-F. Fan and C.-J. Liao. Many-valued modal logic and regular equivalences in weighted social networks. In *Symbolic and Quantitative Approaches to Reasoning with Uncertainty*, pages 194–205. Springer, 2013.
- [33] K. Faust. Triadic configurations in limited choice sociometric networks: Empirical and theoretical results. *Social Networks*, pages 273–282, 2008.
- [34] S. Field, K. A. Frank, K. Schiller, C. Riegle-Crumb, and C. Muller. Identifying positions from affiliation networks: Preserving the duality of people and events. *Social Networks*, 28(2):97–123, 2006.
- [35] N. E. Friedkin and E. C. Johnsen. Social positions in influence networks. *Social Networks*, 19(3):209–222, 1997.
- [36] K. Fujimoto and T. W. Valente. Social network influences on adolescent substance use: Disentangling structural equivalence from cohesion. *Social Science & Medicine*, 74(12):1952–1960, 2012.
- [37] A. P. Gasch and M. B. Eisen. Exploring the conditional coregulation of yeast gene expression through fuzzy k-means clustering. *Genome Biol*, 3(11), October 2002.
- [38] G. B. Giannakis, F. Bach, R. Cendrillon, M. Mahoney, and J. Neville. Signal processing for big data, 2014.
- [39] E. Gleave, H. T. Welser, T. M. Lento, and M. A. Smith. A conceptual and operational definition of ‘social role’ in online community. In *42nd Hawaii International Conference on System Sciences*, pages 1–11, 2009.
- [40] B. Gliwa, A. Zygmunt, and J. Koźlak. Analysis of roles and groups in blogosphere. In *Proc. of the 8th Intl. Conference on Computer Recognition Systems*, pages 299–308, 2013.
- [41] S. A. Golder and J. Donath. Social roles in electronic communities. *Internet Research*, 5:19–22, 2004.
- [42] M. C. González, H. J. Herrmann, J. Kertész, and T. Vicsek. Community structure and ethnic preferences in school friendship networks. *Physica A: Statistical mechanics and its applications*, 379(1):307–316, 2007.
- [43] R. A. Hanneman and M. Riddle. Introduction to social network methods, 2005.

- [44] J. Hautz, K. Hutter, J. Fuller, K. Matzler, and M. Rieger. How to establish an online innovation community? the role of users and their innovative content. In *Hawaii Intl. Conference on System Sciences*, pages 1–11, 2010.
- [45] X. He, H. Zha, C. H. Ding, and H. D. Simon. Web document clustering using hyperlink structures. *Computational Statistics & Data Analysis*, 41(1):19–45, 2002.
- [46] P. Holland and S. Leinhardt. An Omnibus Test for Social Structure Using Triads. *Sociological Methods and Research*, 7:227–256, 1978.
- [47] J. E. Jackson. *A User’s Guide to Principal Components*. John Wiley & Sons, 2004.
- [48] M. Jamali and H. Abolhassani. Different aspects of social network analysis. In *Intl. Conference on Web Intelligence*, pages 66–72. IEEE, 2006.
- [49] R. Jin, V. E. Lee, and H. Hong. Axiomatic ranking of network role similarity. In *Proc. of Intl. Conference on Knowledge Discovery and Data Mining*, pages 922–930. ACM, 2011.
- [50] M. Jung and M. Choi. A mechanism of institutional isomorphism in referral networks among hospitals in seoul, south korea. *The health care manager*, 29(2):133–146, 2010.
- [51] D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the spread of influence through a social network. In *Proc. of 9th ACM Intl. Conference on Knowledge Discovery and Data Mining*, pages 137–146, 2003.
- [52] S. Khot. Improved inapproximability results for maxclique, chromatic number and approximate graph coloring. In *Proc. of IEEE Symposium on Foundations of Computer Science*, pages 600–609, 2001.
- [53] E. L. Kick, L. A. McKinney, S. McDonald, and A. Jorgenson. A multiple-network analysis of the world system of nations, 1995–1999. *Sage handbook of social network analysis. Thousand Oaks, CA: Sage Publications*, pages 311–27, 2011.
- [54] R. Kumar, J. Novak, and A. Tomkins. Structure and evolution of online social networks. In *Link mining: models, algorithms, and applications*, pages 337–357. Springer, 2010.
- [55] H. Kwak, C. Lee, H. Park, and S. Moon. What is Twitter, a Social Network or a News Media? In *Proc. of 19th Intl. World Wide Web Conference*, pages 591–600, 2010.
- [56] V. Labatut and J.-M. Balasque. Detection and interpretation of communities in complex networks: Practical methods and application. In *Computational Social Networks*, pages 81–113. Springer, 2012.

- [57] A. Lampinen, S. Tamminen, and A. Oulasvirta. All my people right here, right now: management of group co-presence on a social networking site. In *Proc. of ACM Intl. Conference on Supporting Group Work*, pages 281–290. ACM, 2009.
- [58] D. Laniado, R. Tasso, Y. Volkovich, and A. Kaltenbrunner. When the wikipedians talk: Network and tree structure of wikipedia discussion pages. In *Intl. Conference on Weblogs and Social Media*, 2011.
- [59] J. Leskovec and C. Faloutsos. Sampling from Large Graphs. In *Proc. of ACM Conference on Knowledge Discovery and Data Mining*, 2006.
- [60] R. Li, S. Wang, H. Deng, R. Wang, and K. C.-C. Chang. Towards social user profiling: unified and discriminative influence model for inferring home locations. In *Proc. of Intl. Conference on Knowledge discovery and data mining*, pages 1023–1031. ACM, 2012.
- [61] L. Lipsky. *Queueing Theory: A Linear Algebraic Approach*. Springer-Verlag, 2nd edition, 2009.
- [62] F. Lorrain and H. C. White. Structural equivalence of individuals in social networks. *The Journal of mathematical sociology*, 1(1):49–80, 1971.
- [63] R. Malcolm, C. Morrison, T. Grandison, S. Thorpe, K. Christie, A. Wallace, D. Green, J. Jarrett, and A. Campbell. Increasing the accessibility to big data systems via a common services api. In *IEEE International Conference on Big Data*, pages 883–892. IEEE, 2014.
- [64] S. Maniu, T. Abdessalem, and B. Cautis. Casting a web of trust over wikipedia: an interaction-based approach. In *Proceedings of the 20th international conference companion on World wide web*, pages 87–88. ACM, 2011.
- [65] F. T. McAndrew and H. S. Jeong. Who does what on facebook? age, sex, and relationship status as predictors of facebook use. *Computers in Human Behavior*, 28(6):2359–2365, 2012.
- [66] O. Medelyan, I. H. Witten, and D. Milne. Topic indexing with wikipedia. In *Proceedings of the AAAI WikiAI workshop*, pages 19–24, 2008.
- [67] A. Mislove, M. Marcon, K. Gummadi, P. Druschel, and B. Bhattacharjee. Measurement and Analysis of Online Social Networks. In *Proc. of the ACM Internet Measurement Conference*, 2007.
- [68] J. Moody. Matrix methods for calculating the triad census. *Social Networks*, pages 291–299, 1998.

- [69] D. Morselli, D. Spini, and T. Devos. Human values and trust in institutions across countries: A multilevel test of schwartzs hypothesis of structural equivalence. *Survey Research Methods*, 6(1):49–60, 2012.
- [70] M. Newman. *Networks: an introduction*. Oxford University Press, 2010.
- [71] R. D. Nolker and L. Zhou. Social computing and weighting to identify member roles in online communities. In *Web Intelligence, 2005. Proceedings. The 2005 IEEE/WIC/ACM International Conference on*, pages 87–93, 2005.
- [72] T. Opsahl and P. Panzarasa. Clustering in weighted networks. *Social networks*, 31(2):155–163, 2009.
- [73] F. Pallotti and A. Lomi. Network influence and organizational performance: The effects of tie strength and structural equivalence. *European Management Journal*, 29(5):389–403, 2011.
- [74] T. A. Pempek, Y. A. Yermolayeva, and S. L. Calvert. College students’ social networking experiences on facebook. *Journal of Applied Developmental Psychology*, 30(3):227–238, 2009.
- [75] S. M. Radil, C. Flint, and G. E. Tita. Spatializing social networks: Using social network analysis to investigate geographies of gang rivalry, territoriality, and violence in los angeles. *Annals of the Association of American Geographers*, 100(2):307–326, 2010.
- [76] R. B. Rothenberg, J. J. Potterat, D. E. Woodhouse, S. Q. Muth, W. W. Darrow, and A. S. Klov Dahl. Social network dynamics and hiv transmission. *Aids*, 12(12):1529–1536, 1998.
- [77] M. Rowe, M. Fernandez, S. Angeletou, and H. Alani. Community analysis through semantic rules and role composition derivation. *Web Semantics: Science, Services and Agents on the World Wide Web*, 18(1):31–47, 2013.
- [78] J. Scott and P. J. Carrington. *The SAGE handbook of social network analysis*. SAGE publications, 2011.
- [79] G. Simmel and K. H. Wolff. *The Sociology of Georg Simmel*. Macmillan Publishing Co., 1950.
- [80] J. B. Singer. User-generated visibility: Secondary gatekeeping in a shared media space. *New Media & Society*, 16(1):55–73, 2014.
- [81] M. M. Skeels and J. Grudin. When social networks cross boundaries: a case study of workplace use of facebook and linkedin. In *Proc. of ACM Intl. Conference on Supporting Group Work*, pages 95–104. ACM, 2009.

- [82] D. A. Smith and D. R. White. Structure and dynamics of the global economy: Network analysis of international trade 1965-1980. *Social Forces*, 70(4):857–893, 1992.
- [83] P.-N. Tan, M. Steinbach, and V. Kumar. *Introduction to Data Mining*. Addison-Wesley, 2006.
- [84] J. Tang, J. Sun, C. Wang, and Z. Yang. Social influence analysis in large-scale networks. In *Proc. of ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 807–816. ACM, 2009.
- [85] A. Tselykh and G. Veselov. Positional analysis and mapping of scientific networks. *World Applied Sciences Journal*, 27(12):1625–1629, 2013.
- [86] J. Ugander, B. Karrer, L. Backstrom, and C. Marlow. The Anatomy of the Facebook Social Graph. Technical report, arXiv:1111.4503v1 [cs.SI], 2001.
- [87] B. Viswanath, A. Mislove, M. Cha, and K. Gummadi. On the Evolution of User Interaction in Facebook. In *Proc. of 2nd ACM Workshop on Online Social Networks*, 2009.
- [88] H. T. Vu. The online audience as gatekeeper: The influence of reader metrics on news editorial selection. *Journalism*, 15(8):1094–1110, 2014.
- [89] S. Wasserman and K. Faust. *Social network analysis: Methods and applications*, volume 8. Cambridge university press, 1994.
- [90] S. S. Wasserman. Random Directed Graph Distributions and the Triad Census in Social Networks. *Journal of Mathematical Sociology*, 5:61–86, 1977.
- [91] B. Wellman. Structural analysis: From method and metaphor to theory and substance. *Contemporary Studies in Sociology*, 15:19–61, 1997.
- [92] H. T. Welser, D. Cosley, G. Kossinets, A. Lin, F. Dokshin, G. Gay, and M. Smith. Finding social roles in wikipedia. In *Proc. of ACM iConference*, pages 122–129, 2011.
- [93] A. J. White, J. Chan, C. Hayes, and B. Murphy. Mixed membership models for exploring user roles in online fora. In *Intl. Conference on Weblogs and Social Media*, 2012.
- [94] H. C. White. Varieties of markets. *Contemporary Studies in Sociology*, 15:226–260, 1997.
- [95] H. C. White, S. A. Boorman, and R. L. Breiger. Social structure from multiple networks. i. blockmodels of roles and positions. *American journal of sociology*, pages 730–780, 1976.



- [96] C. Wilson, B. Boe, A. Sala, K. P. Puttaswamy, and B. Y. Zhao. User interactions in social networks and their implications. In *Proceedings of the 4th ACM European conference on Computer systems*, pages 205–218. ACM, 2009.
- [97] C. Wilson, A. Sala, K. P. Puttaswamy, and B. Y. Zhao. Beyond social graphs: User interactions in online social networks and their implications. *ACM Transactions on the Web*, 6(4):17, 2012.
- [98] A. Zaheer and G. G. Bell. Benefiting from network position: firm capabilities, structural holes, and performance. *Strategic management journal*, 26(9):809–825, 2005.
- [99] Y. Zhang, M. Chen, S. Mao, L. Hu, and V. C. Leung. Cap: community activity prediction based on big data analysis. *IEEE Network*, 28(4):52–57, 2014.
- [100] D. Zhong and H. Zhang. Clustering methods for video browsing and annotation. Technical report, In SPIE Conference on Storage and Retrieval for Image and Video Databases, 1997.
- [101] M. Zhou and C.-u. Park. The cohesion effect of structural equivalence on global bilateral trade, 1948–2000. *International Sociology*, 27(4):502–523, 2012.
- [102] T. Zhu, B. Wang, B. Wu, and C. Zhu. Role defining using behavior-based clustering in telecommunication network. *Expert Systems with Applications*, 38(4):3902–3908, 2011.